Architectural Support for Multiprocessing on Pentium® II Xeon™ Processor Based Workstations

June 1998

Order Number: 283036-001

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

DISCLAIMERS AND RESTRICTIONS

Performance report notice

INTRODUCTION

The steady rise in the performance of Intel microprocessors has made it possible to build powerful Intel Architecture workstations that compete with and outpace RISC alternatives in performance while preserving the cost advantage associated with Intel-based systems. With the Pentium® II Xeon™ processor, dual-processor platforms provide substantial performance benefits.

Platforms based on the Pentium II Xeon processor incorporate a wide range of architectural features needed to support multithreaded and multitasking workstation workloads. In this document, we briefly describe these features, present measurements indicating the high levels of performance that can be achieved, and show that multiprocessing capabilities add only a small increase to overall system cost.

OVERVIEW OF THE INTEL ARCHITECTURE FOR WORKSTATIONS

A block diagram representing a typical Intel Architecture workstation is depicted in Figure 1. The platform is organized as a shared-memory symmetric multiprocessor (SMP), with a high-bandwidth bus being used to connect the processors to memory and I/O.

While most current implementations are based on a dual-processor configuration, the architecture can easily be extended to support more processors. Indeed, one important feature of Intel Architecture workstations is their support for easy upgradability, where processors can be added/removed with no changes to the underlying platform hardware. The MP kernel of Windows* NT automatically configures itself to utilize all the processing resources available at system start-up time.

In order to support parallel computing, the architecture must provide mechanisms for process communication and synchronization. Communication is readily supported by reserving part of the address space to store shared data, and allowing processes to read from and write into this shared space. Synchronization, on the other hand, is supported by providing atomic memory read-modify-write instructions, which are used by the operating system to construct a variety of useful high-level synchronization constructs, including locks and barriers.

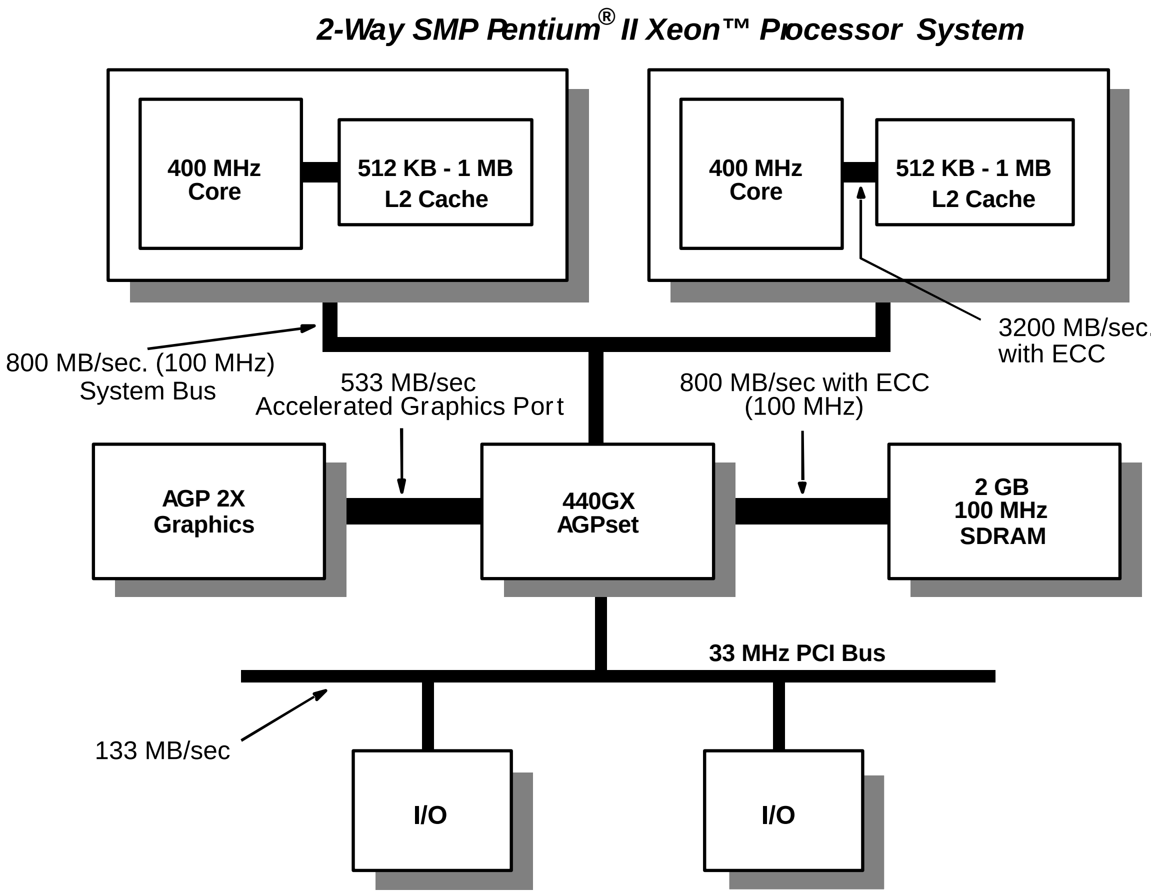

Figure 1 Key architectural features of multiprocessor workstations built on the Intel Pentium® II Xeon™ processor and 440GX AGPset

While resource allocation and process scheduling are also functions of the operating system, support for these functions at the architectural level improves the efficiency of operating system functions, which can improve application performance. For example, the architecture provides extensive interrupt handling capabilities which can be utilized by the operating system to perform fast context switching. Priority levels can also be assigned to various interrupts, which helps in implementing process priorities and incorporating these priorities in making scheduling decisions.

The presence of multiple superscalar processors in the system substantially increases the demand for main memory bandwidth. However, the Pentium II Xeon processor and its 440GX AGPset incorporate a wide range of features intended to address this demand and prevent the bus-based processor/memory interconnection from becoming a performance bottleneck. In short, the Pentium II Xeon processor based system is a great deal more than just a higher speed processor – it is a balanced system. We will now describe these features in some detail

PROCESSOR FEATURES

The Pentium II Xeon Processor is designed for glueless multiprocessing. Up to four processors can be connected to the system bus and the Advanced Programmable Interrupt Controller (APIC) bus without any external glue logic. Such a system allows for normal MP behavior without any special ASICs. The processors implement an invalidation based snoopy cache coherence protocol. The system bus contains all the signals necessary to implement this protocol, as do the 450NX PCIset and the 440GX AGPset, which were designed to work with the Pentium II Xeon Processor. The processors arbitrate amongst themselves in order to gain

ownership of the bus. The arbitration mechanism is distributed amongst the processors and allows bus ownership to be exchanged very rapidly even with the latched bus protocol. No external logic is required for this purpose.

In addition to providing an integrated interrupt bus that allows for efficient inter-processor interrupts, the interrupt mechanism supports an MP boot protocol. This protocol allows an MP system to boot up without any special BIOS or OS support.

The Pentium® II Xeon™ processor makes extensive use of high speed caches to reduce access delays to the shared main memory. The processor uses a 2-level caching structure. Two 16 KB 2 -way set associative L1 caches are built into the processor core. The L1 caches are separated into instruction and data caches, allowing instruction fetch and data access to proceed in parallel. In addition, each processor has a dedicated 4 - way set associative, shared instruction and data L2 cache. This L2 cache is closely coupled to the processor through a dedicated 64 bit bache bus which runs at the same frequency as the processor core. This feature guarantees that data from the L2 cache arrives at the CPU in the shortest possible time. Three different L2 cache sizes are available (512KB, 1MB, and 2MB) for Pentium II Xeon systems to accommodate the needs of a variety of WS workloads. Table 1 shows cache misses per thousand instructions retired (MPTI) for each of the three L2 cache sizes. A lower number indicates more cache data is reused, freeing the main memory bus for other activity.

Benchmark	L1 MPTI	L2 MPTI 512 KB	L2 MPTI 1 MB	L2 MPTI 2 MB
Nastran lgamd	57.1	34.0	25.7	21.0
Ansys ttnb-pm	36.6	14.8	12.2	10.1
Pro/E bench97	9.0	1.5	1.0	0.8
Viewperf CDRS	15.3	6.5	2.9	0.8
SPECfp95 wave5	103.2	9.2	6.7	4.2

Table 1 Pentium II Xeon Processor L1 and L2 MPTI for typical workstation benchmarks

Several other cache optimizations have also been incorporated in the L1/L2 cache design. Cache line size (32 bytes) and associativity level (2-way set associative for L1 instruction cache and 4-way for L1 data and L2 caches) are optimized to reduce miss rate. The L2 write policy uses a write-allocate / write-back strategy to further reduce the demand for memory bandwidth. In addition, write combining is used to improve the efficiency of writes to uncached data, such as data in framebuffers. Partial writes can be combined in the write buffer so that they may be carried out using a burst-mode bus transaction.

Another important feature of the L1/L2 caches is that they are non-blocking; they can continue to supply cache hits even when a miss is pending. This reduces the occurrence of processor pipeline stalls due to cache misses, which is particularly important for superscalar processors. It also enables the cache to hide the latency of write misses by allowing subsequent non-dependent reads to move forward. This essentially implements a memory consistency model called “processor consistency”, which improves performance while maintaining a simple parallel programming model. In a processor consistent system, all loads are performed in order and all stores are performed in order. However, loads are allowed to pass non-conflicting stores.

To support caching of shared data in a multiprocessor environment, a special Modified-Exclusive-Shared-Invalid (MESI) protocol is used to maintain cache coherence. With this protocol, a processor must invalidate all other cached copies of a shared line before it is allowed to actually modify it. Cache controllers constantly snoop on the system bus watching for invalidate commands targeted toward locally-cached lines. The cache system is designed to substantially reduce interference of snoop activity with processor data access. As an added

optimization, cache "snarfing" is implemented to allow cache-to-cache transfers of shared data on a read miss, instead of having to fetch the data from the slower main memory.

Intel has enabled ECC on L2 cache on Pentium II Xeon processors, including continuous single-bit error correction and multiple-bit error detection. This is necessary to protect data on the larger L2 cache sizes available with the processor. The processor-level protection of company assets makes it easy for workstation manufacturers to provide error correction on their business critical applications. With protection on both instructions and data, system failures and corrupted data will not occur as a result of single-bit L2 cache errors.

Finally, the system allows caching of synchronization variables, which prevents the processors from consuming a lot of memory bandwidth for busy waiting. Reducing synchronization overhead improves the speedup derived from parallel processing.

HIGH SPEED SYSTEM BUS

An important goal of the Pentium® II Xeon™ processor bus architecture was to make multiprocessing simple and straightforward for the system designer. This goal can be met if the entire support for multiprocessing is contained in the processor. This requires the processor to handle all arbitration, cache coherency and inter-processor communication. The processor implements a synchronous de-multiplexed pipelined bus. The system bus is 64 bits wide, and it operates at 100 MHz. The theoretical peak bus bandwidth is 800 MB/sec. Interprocessor interrupt communication is supported on the processor bus via the Advanced Programmable Interrupt Controller (APIC). Multiprocessing complexity is completely hidden within the processor. The bus protocol defines crisp ordering and cache coherency rules to be followed by all agents on the bus.

The bus operates in several pipelined phases. An agent on the bus initiates a transaction by making a request for the bus. Busagents arbitrate amongst themselves during the **arbitration** phase to determine the owner. The processors arbitrate in a round robin fashion while I/O agent requests intervene at a higher priority. The bus owner places its request on the bus during the **request** phase. All agents check the request for correct parity and can signal any errors in the **error** phase. Multiprocessor cache coherency support is an integral part of the processor bus protocol, which uses the four MESI cache states. The bus agents drive the current cache state on the bus during the **snoop** phase. All bus agents can observe snoop results and complete the resultant cache state transitions within the bus clock. This ensures that all agents will update their internal cache state during a transaction and reflect these results in the next transaction whose snoop phase follows three cycles later. The transaction then completes during a **response** phase, with actual data transfer taking place during the **data** phase. All transactions on the bus are normally completed in order. However, devices that have long latencies can choose to defer a transaction and respond to it later. This allows other subsequent transactions to complete without being held up. Once the long latency device is ready to reply it arbitrates for the bus and completes the transaction that it had deferred. Up to eight transactions can be outstanding on the bus simultaneously, with no more than four from any single processor.

AGPSET FEATURES

A powerful high-end processor is not the only contribution that Intel makes to building high-performance workstation platforms. A new AGPset, called the 440GX AGPset, has been specially designed to support high-speed access to large sizes of main memory and provide the proper interface to AGP and multiple PCI devices.

Accelerated Graphics Port (AGP) is an important innovation that substantially improves the performance of graphics-intensive workstation applications. Instead of using a PCI interface to support high-end graphics cards (which forces the card to share limited PCI bandwidth with other peripherals such as network and SCSI adapters), AGP provides a separate high-bandwidth data pathway to main memory (533 MB/sec versus 133 MB/sec theoretical maximum on PCI). The use of graphics accelerators adds a new dimension to the capabilities of

Intel Architecture workstations since texture data fetching and manipulation by the accelerator can proceed in parallel with other activities being carried out by the processors.

The 440GX supports up to 2 GB of physical main memory, thereby accommodating the needs of a wider variety of multiprocessing applications. Sufficient memory resources are particularly important for multitasking workloads. On another front, two critically important features are provided by the 440GX AGPset to increase the effective memory bandwidth available to uniprocessor as well as multiprocessor applications. The first is support for a significant increase in memory bandwidth. The memory bus is 64 bits wide and accesses Synchronous DRAMs at 100 MHz, providing a maximum theoretical bandwidth of 800 MB/sec.

The second feature involves the techniques used by the AGPset to improve the latency and throughput of access to SDRAM-based main memory. When a new row address is provided to the SDRAM, a new memory "page" is "opened," and subsequent accesses to the same page require only column addresses to be provided; the AGPset capitalizes on this to improve latency. The AGPset also breaks up each memory access into separate stages (RAS/CAS/data) so that successive references can be processed in a pipelined fashion, thus improving throughput.

To further enhance the speed with which main memory can be accessed, the 440GX allows multiple pages to remain open within the SDRAM array. This feature is important in multiprocessor configurations because each processor is likely to access different data structures. Unless multiple pages are kept open, the interleaving of the accesses coming in from the processors will result in frequent page misses, which produces bubbles in the memory access pipeline, thereby significantly curtailing overall memory throughput and slowing overall performance.

Furthermore, to achieve better utilization, the 440GX fully supports split transactions on the 100 MHz system bus and provides buffering for multiple outstanding memory requests. Split transactions mean that each bus access is handled as separate request and response operations, so that the bus is not held for the entire duration of a transaction.

System bus bandwidth utilization varies across different applications and data sets. Table 2 shows the measured system bus bandwidth for several application workloads for a Pentium II Xeon system. Note that the bandwidth is well below the theoretical maximum provided by the system. This headroom assures excellent performance scaling when a second processor is added to the system.

Benchmark	System Bus Bandwidth (MB/sec)
Nastran lgamd	135.7
Ansys slsb-pm	76.9
Pro/E bench97	16.2
Viewperf CDRS	101.1
Verilog beh160	185.5
SPECint95	0.7 to 62.9
SPECfp95	0.3 to 352.9

Table 2 System Bus Bandwidth for typical workstation benchmarks

PERFORMANCE

Having discussed the architectural features intended to support multiprocessing, this section will quantify the high levels of performance that can be achieved on workstations based on the Pentium II Xeon processor. Intel has tested several multithreaded benchmarks provided by workstation software vendors, and constructed several benchmarks to test multitasking performance. Benchmarks were run on systems containing dual Pentium II Xeon processors. A complete system configuration is found in the Appendix. A descriptions of these benchmarks can be found in a companion paper [*Breakthrough Workstation Performance with the Intel® Pentium® II Xeonä Processor, June 1998, Order Number 283034-001*]. The results are shown in Figure 2 and 3.

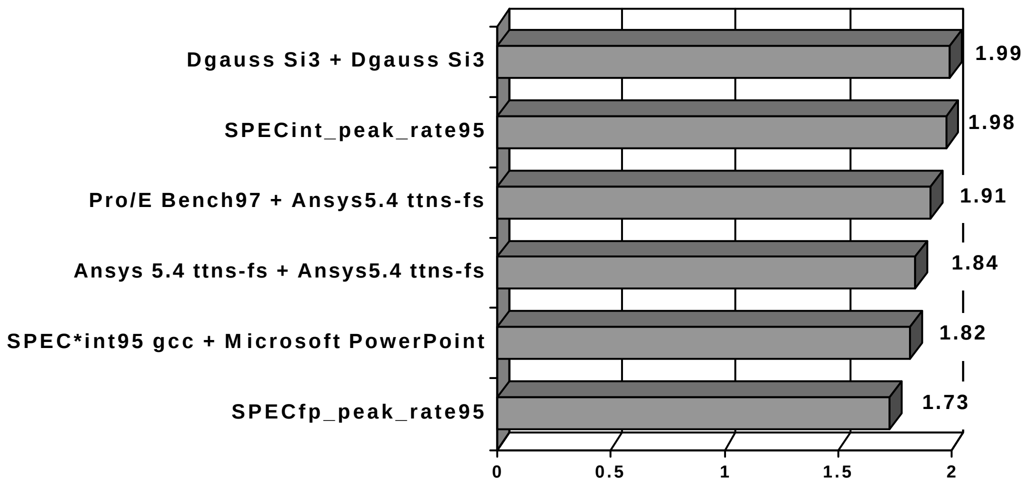

Figure 2 Multi-tasking performance multiplier in six tests

Multi-Tasking Test Case Elapsed Time (seconds)	Uni-Processor	Dual-Processor	Performance Multiplier
Dgauss Si3 + Dgauss Si3	3,485.7	1,752.3	1.99
SPECint_peak_rate95	146.7	290.8	1.98
Pro/E Bench97 + Ansys5.4 ttns-fs	2,946	1,544	1.91
Ansys 5.4 ttns-fs + Ansys5.4 ttns-fs	479.4	260.7	1.84
SPEC*int95 gcc + Microsoft PowerPoint	200.2	110.3	1.82
SPECfp_peak_rate95	119.6	206.8	1.73
Average Performance Multiplier			***1.88***

Table 3 Table summarizing the multi-tasking performance multiplier in six tests

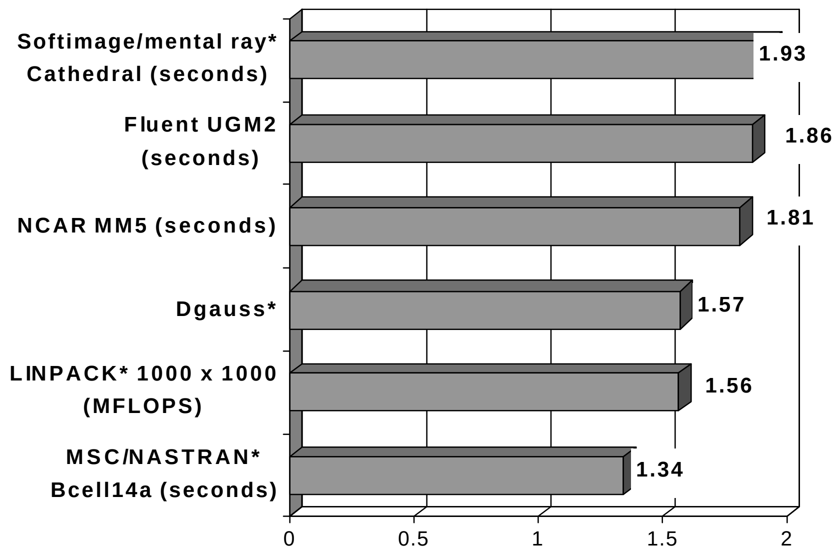

Figure 3 Multi-threaded performance multiplier in six tests

Multi-Threaded Test Case	Uni-Processor	Dual-Processor	Performance Multiplier
MSC/NASTRAN* Bcell14a (seconds)	458.40	341.11	1.34
LINPACK* 1000 x 1000 (MFLOPS)	357.90	229.90	1.56
Dgauss (seconds)	28.66	18.31	1.57
NCAR MM5 (seconds)	5852	3234	1.81
Fluent UGM2 (seconds)	2196	1180	1.86
Softimage/mental ray* Cathedral (seconds)	462.00	239.33	1.93

Table 4 Table summarizing the multi-threaded performance multiplier in six tests

When looking at the performance of multithreaded applications, it is important to keep in mind that the amount of parallelism inherently available in the computation puts an upper bound on the speedup that can be achieved. This, however, is not a limiting factor for multitasking applications. When there is enough parallelism to keep both processors busy, speedup diminishes only when contention for shared computational resources is high enough to cause requests from the two processors to be delayed. Main memory resources are particularly important in this regard; 512 MB of memory was used for all of the multitasking benchmarks to minimize disk I/O.

COST

The relative and absolute results of these tests are remarkable in and of themselves. They are even more compelling when one considers the cost of a dual Pentium II Xeon processor-based workstation and the total cost of ownership. As a rule of thumb, the cost of an additional processor adds approximately 15% to the total system cost. However, the cost is justified considering the performance potential of a dual-processor system as well as the ability to use a single system for running workstation and office productivity applications.

A formula for calculating the time-to-payback for an incremental Intel based-processor is provided in the Intel white paper *"Cost/Performance Benefits of Multi-Tasking on Intel Architecture Multiprocessor Workstations," June 1998, Order Number 283038-001*.

SUMMARY

Workstations based on Intel's Pentium II Xeon processor and its 440GX AGPset provide the full range of architectural features needed to efficiently support multiprocessing. Dual-processing capabilities add little to the overall cost of these systems while at the same time delivering a high level of performance on a variety of multithreaded and multitasking workstation workloads. Software developers have a lot to gain by taking advantage of the multiprocessing features of today's powerful Intel Architecture workstations.

An overview of multithreading is included in a companion white paper titled "*Multi-Threading: Taking Advantage of Intel-Architecture Multiprocessor Workstations," June 1998, Order Number 283037-001*.

APPENDIX A — TEST CONFIGURATIONS

Specification	Intel® Pentium® II Xeon™ Processor
Processor	Pentium II Xeon with 512 KB level 2 cache, 400 MHz clock rate, 100 MHz system bus, ECC on
Mother board	MS440GX with Alpha 5 BIOS and Intel 440GX AGPset (440BX AGPset on Multithreaded tests)
Memory	SDRAM, ECC on
Operating System Configuration	Windows* NT* 4.0 1381 SP3 with NTFS file system on disk

Table 5 Basic configuration used for benchmarking the Intel Pentium® II Xeon™ processor

Benchmark	Memory	Graphics	Disk Subsystem	Networking Card
MSC/Nastran* Softimage*	256 MB	Evans & Sutherland* AccelEclipse* 32 MB, 1.3.09 Driver, Colors=TRUE	Seagate* Cheetah* ST34501W, Adaptec* 2940 UW Controller	(None)
Pro/ENGINEER*	256 MB	E&S AccelEclipse 32 MB, 2.0 Realimage AGP Driver, Colors=TRUE	Seagate Cheetah ST34501W, Adaptec 2940 UW Controller	EtherExpress* Pro/100B
Viewperf	128 MB	E&S AccelEclipse 32 MB, 2.0 Realimage AGP Driver, Colors=TRUE	Seagate Cheetah ST34501W, Adaptec 2940 UW Controller	EtherExpress Pro/100B
SPECint95* SPECfp95*	64 MB	Matrox* Millennium* II, 4MB, 3.11, 4.0.42 Driver, Colors=16	Seagate Cheetah ST34501W, Adaptec 2940 UW Controller	(None)
SPEC_rate_int95* SPEC_rate_fp95*	256 MB	Matrox Millennium II, 4MB, 3.11, 4.0.42 Driver, Colors=16	Seagate Cheetah ST34501W, Adaptec 2940 UW Controller	(None)
ANSYS* 5.4 Linpack* 1000x1000	512 MB	E&S AccelEclipse 32 MB, 2.0 Realimage AGP Driver, Colors=TRUE	Seagate Cheetah ST34501W, Adaptec 2940 UW Controller	(None)
Multi-tasking workloads (except below)	512 MB	E&S AccelEclipse 32 MB, 2.0 Realimage AGP Driver, Colors=TRUE	Seagate Cheetah ST34501W, Adaptec 2940 UW Controller	EtherExpress Pro/100B
SPECint95 gcc and Microsoft* PowerPoint* workload	512 MB	Matrox Millennium II, 4MB, 3.11, 4.0.42 Driver, Colors=16	Seagate Cheetah ST34501W, Adaptec 2940 UW Controller	EtherExpress Pro/100B

Table 6 Specific configurations used for the Intel Pentium® II Xeon™ Processor measurements

APPENDIX B — AVERAGE SYSTEM PRICE

Vendor	Model	Processor Speed	Dual Processor System Price (Est. Street Price)
Major Intel Architecture Workstation Vendor “A”	Dual Processor, 2D Graphics	333 MHz	$7,377
		400 MHz	$7,676
	Dual Processor, 3D Graphics	333 MHz	$8,327
		400 MHz	$8,626
Major Intel Architecture Workstation Vendor “B”	Dual Processor, 2D Graphics	333 MHz	$7,679
		350 MHz	$7,925
		400 MHz	$8.520
	Dual Processor, 3D Graphics	400 MHz	$9,438
Major Intel Architecture Workstation Vendor “C”	Dual Processor, 2D Graphics	333 MHz	$8,131
	Dual Processor, 3D Graphics	333 MHz	$12,216

Table 7 Estimated Street Prices for Dual Processor Workstations, May 1998

Note: Prices are approximate. Intel Architecture vendor pricing is for workstations based on the Pentium II processor; pricing for workstations based on the Pentium II Xeon processor was not available at the time of publication. Data was gathered from public web sites, workstation manufacturers and/or their distributors. All configurations include 512 MB memory, 21-inch monitor, local hard drive and a network card. Some configurations include sound systems (card, speakers). 2D and 3D graphics configurations and capabilities vary.

Website References

- **Intel**
  http://www.intel.com
- **SPEC**
  http://www.specbench.org/osg/cpu95/results/results.html
- **BAPCo SYSmarkNT4.0**
  http://www.bapco.com/ntrslts.htm
- **ANSYS**
  http://www.ansys.com/VisitAnsys/Pubs/AnsysNews/Q4/page20/page20_partn.html
- **Dgauss: Oxford Molecular Group**
  http://www.oxmol.com
- **Pro/ENGINEER**
  http://www.proe.com
- **Viewperf**
  http://www.specbench.org/gpc/opc
- **MSC/NASTRAN**
  http://134.134.214.1/procs/ppro/wrkstn/brief/msc.htm
- **Fluent**
  http://www.fluent.com
  http://www.kai.com
- **MM5**
  http://www.mm5.ucar.edu/mm5
- **LINPACK**
  http://www.netlib.org/benchmark/linpackjava/
  http://www.extreme.indiana.edu/hpjava/papers/loop/node13.html
  http://www.netlib.org/benchmark/top500/lists/linpack.html

- **Sun Microsystems, Inc.**
  http://www.sun.com/
- **Hewlett-Packard Co.**
  http://www.hp.com/
- **International Business Machines**
  http://www.ibm.com/
- **Dell Computer Corp.**
  http://www.dell.comm/
- **Silicon Graphics**
  http://www.sgi.comm/
- **Gateway 2000**
  http://www.gateway2000.comm/
- **NEC**
  http://www.nec.comm/
- **Siemens Nixdorf**
  http://www.sni.comm/
- **Compaq**
  http://www.compaq.comm/
- **Fujitsu**
  http://www.fujitsu.comm/
- **Intergraph**
  http://www.intergraph.comm/